DOCUMENT RESUME

ED 349 323 TM 018 947

AUTHOR Hyers, Albert D.; Anderson, Paul S.
TITLE Methods and Materials for Geography Education Improvement through Innovative, Machine-Scored Assessment.
PUB DATE 91
NOTE 11p.; Paper presented at the Conference of the New England-St. Lawrence Valley Geographical Society (Montreal, Quebec, Canada, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Students; Comparative Testing; *Computer Assisted Testing; Difficulty Level; Educational Assessment; Educational Improvement; Educational Innovation; *Geography; Higher Education; *Instructional Materials; Multiple Choice Tests; *Scoring; Teaching Methods; Test Format; Test Items; Test Reliability
IDENTIFIERS Item Discrimination (Tests)

ABSTRACT

Using matched pairs of geography questions, a new testing method for machine-scored fill-in-the-blank, multiple-digit testing (MDT) questions was compared to the traditional multiple-choice (MC) style. Data were from 118 matched or parallel test items for 4 tests from 764 college students of geography. The new method produced superior results when measured by three widely accepted education statistics: (1) item difficulty; (2) item discrimination; and (3) test reliability (Kudor-Richardson Formula 20). MDT equivalents of world geography MC questions were more difficult and had 14% better discrimination. Map location questions were much easier than concept questions regardless of style. Furthermore, in contrast to the traditional MC format, the MDT method more clearly identified marginal examinees as indicated by standardized tests and especially by cumulative grade point averages. The analysis suggests that the MDT method fosters increased learning of geography because it discourages guessing and recognition as examination strategies. Geography teachers can effectively and easily use this method. Two tables and seven figures illustrate the discussion. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
ALBERT D. HYERS
PAUL S. ANDERSON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED349838

METHODS AND MATERIALS FOR GEOGRAPHY EDUCATION IMPROVEMENT THROUGH INNOVATIVE, MACHINE-SCORED ASSESSMENT

A Paper Accepted for Publication in the 1991 *Proceedings* of the
New England-St. Lawrence Valley Geographical Society
and Presented at the 1991 Conference in Montreal

by

Albert D. Hyers
Department of History, Political Science and Geography
North Adams State College
North Adams, MA 01247

and

Paul S. Anderson
Department of Geography-Geology
Illinois State University
Normal, IL 61761

TM018947

BEST COPY AVAILABLE

ABSTRACT

Using matched pairs of geography questions, we compare a new testing method for machine-scored "fill-in-the-blank" (MDT) questions to the traditional multiple choice (MC) style. The new test-ing method produces superior results when measured by three widely accepted education statistics: item difficulty, item discrimination, and test reliability (KR20). MDT equivalents of world regional geography MC questions are more difficult (averaging 6.4 points less on a 100 point test) and have 14% better discrimination. "Map location" questions were much easier than "concept" questions regardless of style. Furthermore, in contrast to a traditional multiple choice format, the MDT method more clearly identified marginal examinees as indicated by standardized tests (A.C.T. or S.A.T.) and especially by cumulative grade point averages. Our analysis suggests that the MDT method fosters increased learning of geography because it discourages guessing and recognition as exam strategies. Traditional grading scales stressing such thresholds as 90% for an "A" are detrimental to assessment because maximum discrimination power occurs in items closer to the 50% level of difficulty. Teachers of geography can effectively and easily employ this method. Relevant materials for several geography courses are available from the authors.

INTRODUCTION

The Problem

The most common objective examination style used in large enrollment, introductory courses is multiple choice. One reason is its long tradition, reinforced even now by sample questions provided in the instructor manuals for popular textbooks. Another reason is that teachers can grade such exams easily with computerized scanning and scoring. Although multiple choice exams have proven reliable, we believe that there are better machine-scored alternatives to assess learning.

In this paper, we describe and analyze an alternative that is essentially a machine-scored equivalent of the traditional "fill-in-the-blank" testing method that requires time consuming manual grading. This new method that we describe here utilizes a separate alphabetized and consecutively numbered long list of up to 1,000 alternatives which we call an "answer bank." Teachers can use a specific answer bank for several course examinations. Beside each entry on an answer bank is a multiple-digit code number (usually three digits) that examinees mark on a specially designed answer sheet or card that an optical mark reader (OMR) can scan. Anderson (1990) describes this method, which also allows for written answers and other unique responses that can be partially machine-scored. One fundamental aspect of such a long list is that it strongly discourages the examinees from reading the entire list to find or recognize the correct answer. Instead, the examinee uses the alphabetized list to look up an answer that is already known or derived intellectually and encodes the numerical response on the answer form. Based on many dozens of examinations with thousands of students, we believe that this format is very nearly equivalent to fill-in-the-blank style questions, but with the advantages of machine scoring. We refer to this method as either the "answer bank" or "MDT" (multiple-digit) style.

The problem under investigation is whether the MDT method offers an improvement of assessment relative to MC (multiple choice). A few studies have already investigated comparisons between the MDT and MC formats (Anderson 1987; Brailovsky et al. 1988; Veloski et al. 1988). Additionally, both of us have observed consistent and noticeable differences between the two formats over several years of writing and evaluating examinations. This paper expands on our recent work (Anderson and Hyers 1991) involving five data sets (including both geography and family practice medicine) from three institutions of higher education. Our purpose here is to elaborate more completely on the differences between the MDT and MC question styles and to draw attention to relationships with examinee "ability." We hypothesize that the MDT method is "better" than MC in terms of established indices of assessment and that MDT should better differentiate students according to "ability."

Methods and Data

Using data from classroom tests, we calculated three well known assessment statistics in our comparison of MDT and MC styles: item difficulty, item discrimination, and the KR20 test reliability coefficient.

(1) Item difficulty is the proportion of examinees answering a question correctly. For example, values of 1.000 and 0.500 mean that all (100%) and half (50%) answered correctly, respectively. The mean of all difficulty indices for an exam, when expressed as a percentage, is equal to the mean test grade.

(2) Item discrimination (or validity) is a measure of how well a question can differentiate among the examinees. In our work, we used the following expression to calculate this index:

$$(CH/NH) - (CL/NL)$$

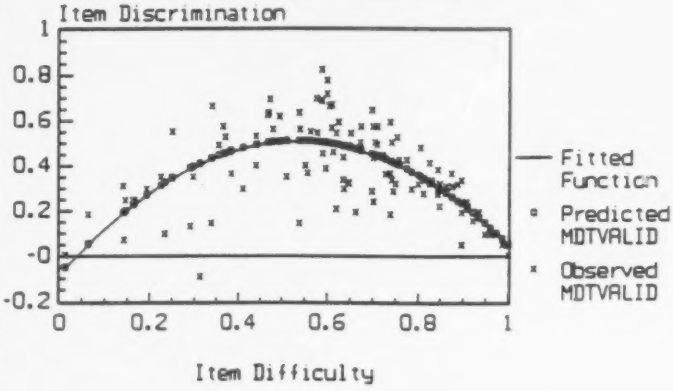
where CH and CL are the number of examinees answering correctly in the highest 27% and lowest 27% of the group, respectively, and NH and NL are the total number of examinees in the respective groups. This method yields values that can range from -1 to +1. Negative values indicate items that the lowest scoring examinees answered better than students with the highest scores; such items are not "valid."

The relationship between difficulty and discrimination is not linear. Figure 1 shows that items with difficulty indices in the 0.3 to 0.7 range are best able to discriminate, and that both the very easy and very difficult items have poor discrimination. The corresponding graph for MC is quite similar but the curve is steeper on both sides because five-choice MC questions have a difficulty index restricted to 0.2 to 1.0. The position of the curve's highest point raises an important issue about common practices in testing: the better exams (at least in terms of being able to discriminate) would produce mainly what many perceive to be failing grades according to widely used grading scales in higher and secondary education. This would occur because many teachers base grades on a scale with thresholds near 70%, 80% and 90% for C, B, and A, respectively. Yet tests with scores that high have lower discrimination power (Fig. 1). It seems that students and teachers alike regard scales like this as "sacred" and that departures from it are somehow undesirable. We note that although these high thresholds are common practice for classroom

assessment, designers of standardized tests (such as the A.C.T. or S.A.T.) try to achieve a mean score around 50%.

(3) Test reliability: The Kudor-Richardson Formula 20 (KR20) is an estimate of the expected correlation between one test and a hypothetical alternative version containing the same number of items. The number of items in the examination, the number of examinees, and their range of "ability" strongly influence KR20 values. Comparisons between data sets require care in interpretation, but observed trends to date are in a direction favorable to the MDT method. Because the basis for the KR20 coefficient is the entire test, rather than individual items, we are currently building a larger data set. In our prior study (Anderson and Hyers 1991), we found that (a) MDT equivalents of multiple choice questions were statistically more difficult (averaging 9.9 points lower on a 100 point test) and (b) better at discrimination (by 17%) and that (c) the MDT format usually had a higher test reliability coefficient (KR20) than MC. That work revealed large variations in the subsets of geography questions and led us to look more closely at the nature of machine-scored geography examinations.

Figure 1. Relationship between Item Difficulty and Item Discrimination for MDT Answer Bank Questions



To better analyze these variations, our present work focuses on the largest geography data set (World Regional Geography). We were able to categorize all questions in that set into three groups: map location, place recall, and concept. For map location items, examinees had to refer to specific places on maps accompanying the exams (e.g., "Location 36 on map J is what country?"), but they did not refer to such a map for place recall items (e.g., "What is the capital of Illinois?"). Concept items included all other questions (e.g., "Which religion has a pilgrimage called the hajj?"). We stress that all the exams used in this analysis were written before we created these categories (as much as 7 years earlier). We occasionally refer to the smaller physical geography data set of concept questions from North Adams State College to help verify the generality of our findings. We applied nonparametric procedures when a data subset did not have a normal distribution.

We analyzed 118 matched or parallel test items from four exams (WG84, WG87, WG90, AND WG91) given between 1984 and 1991. In these experiments, MDT items were restatements of five-choice MC items. The WG84 subset of 158 students involved four course sections from three instructors, and each section answered either MDT or MC. The WG87 subset of 155 students had two random groups answering alternating styles through the semester and two control groups answering either MDT or MC versions exclusively. The WG90 and WG91 subsets of 160 and 291 students, respectively, had examinees answering both styles, but they did not see the MC versions until after handing in the MDT answers (this was also the method used for the PG90 subset of 46 students).

FINDINGS

Item Analysis: Difficulty

As a first step in comparing the MDT and MC styles, we arbitrarily used equal interval difficulty thresholds of 0.33 and 0.67 to classify each question as hard, intermediate, or easy. Most map location questions were in the easy category (78% and 91% for MDT and MC, respectively). The same was true for place recall questions (69% and 75%, respectively) but with so few items (N = 16) we eliminated them from most of our analysis. Concept questions were more difficult, regardless of question style. The MDT style had 16%, 61%, and 23% in the easy, intermediate, and difficult categories, respectively, while the corresponding values for MC were 44%, 37%, and 19%. Thus we observe that the MDT questions as a group were more difficult than MC.

Table 1. Mean difficulty and discrimination indices for selected data sets.

DATA GROUP	N	DIFFICULTY			DISCRIMINATION			
		MDT - MC =		CHANGE	MDT - MC =		CHANGE	CHANGE (%)
Map Location	45	.769	.830	-.061	.384	.315	.069	+22%
Place Recall	16	.658	.724	-.066	.378	.356	.022'	+6%
Concept	57	.500	.566	-.066	.376	.342	.035'	+10%
World Geog.	118	.624	.688	-.064	.380	.334	.046	+14%
Phy. Geog.	13	.478	.686	-.207	.519	.365	.154'	+42%

Notes: "% Change" is from MC to MDT. All P-values for the MDT-MC difference ("Change") are less than .01 except that ' means that P is between 0.01 and 0.02 and ' means P exceeds 0.05.

Table 1 shows that there are large differences in difficulty between the two test styles ($P < .01$) and among the three question categories ($P < .001$). Considering these differences in terms of a 100-point test, the MDT is more difficult than the MC by 6-7 points for all three question categories. This is understandable given that successful guessing and response recognition are almost impossible with sufficiently long MDT answer banks.

We have found that converting test questions from MC to MDT will definitely result in lower scores. These differences in "points" are large relative to the standard 10 point spread in grade categories of the "60-70-80-90" scale. Instructors should adjust grade scales accordingly; the use of a "curve" would be appropriate for teachers starting to use the MDT method. On the other hand, if educators want to require greater "rigor" in a course, conversion to the MDT style of questions will require greater effort by students who want a specific grade (e.g., an A at the 90% level). We contend that using a curve and/or increasing rigor are both desirable.

Item Analysis: Discrimination

The next question we consider is how these lower scores would be an overall "improvement." Table 1 shows that differences in discrimination among the map location, place recall, and concept categories were negligible within each test style. But between test styles, in all cases the discrimination for MDT questions exceeded MC values. However, these differences were statistically significant for only map location ($P < .002$) and concept ($P < .017$) categories. The amount of percent increase (Table 1) for World Regional Geography (averaging 14% but with a maximum of 22%) appears to be a major improvement. This result points out the need to further investigate variations in item discrimination.

These findings show that the reduction in scores accompanying a conversion from MC to MDT results in better discrimination power. One reason that MDT questions have higher discrimination is because they are more difficult and closer to the central range of difficulty as indicated on Figure 1. Our analysis of covariance to control for the effects of item difficulty found for some of the data sets (e.g., intermediate difficulty concept items) that a notable part of the superiority of the MDT method for item discrimination was independent of difficulty differences. Yet similar analyses of covariance for the other, generally smaller, subsets were not statistically significant. Further research should clarify this issue.

We believe that this improvement in discrimination adds further justification for the lower scores and makes an MDT test better than an equivalent (parallel) set of MC questions. This is especially true if one of the purposes of testing is to differentiate levels of learning. Also, because of the lower discrimination of the MC equivalents of MDT questions, we prefer not to use MC (or true-false) for assessing basic or core knowledge where we are expecting high levels of competency. For challenges to students, and for improved discrimination, we prefer to ask some questions in a style other than MC. Of the available styles (e.g., fill-in-the-blank), only the MDT format can be machine-scored.

Regression of Difficulty and Discrimination Indices

We used a least-squares regression to study the relationship between MDT and MC item difficulty and discrimination. Typical examples appear in Figure 2 (for map location items) and

Table 2. Predicted values of MC difficulty and discrimination given respective MDT values of 0.0, 0.5, and 1.0.

DATA GROUP	CATEGORY	DIFFICULTY			DISCRIMINATION		
		0.0	0.5	1.0	0.0	0.5	1.0
World Geography	Map	.448	.697	.945	.043	.398	.752
World Geography	Concept	.139	.566	.994	.128	.412	.696
World Geography	All	.186	.588	.991	.104	.407	.709
Physical Geog.	Concept	.284	.704	1.123	.074	.355	.635

Note: Only Physical Geography correlations (due to small N) have $P > .0001$; $P = .03$ for difficulty and .15 for discrimination.

Figure 3 (for concept items) and summary data are in Table 2. Both of these and the other regressions that we calculated are linear and statistically significant. Question style, as expected, was almost irrelevant for the easiest items because nearly everyone answered correctly with either method. An important observation is that the difference between independent (observed) MDT and dependent (predicted) MC values is greatest for the harder items but much less for the easier ones. Theoretically, if no one knew the answer for an item in either style, the difficulty indices would be 0.001 and 0.2, respectively, for a thousand term MDT answer bank question and a five-choice MC question. In fact, the y-intercept for the WG91 data set of 48 questions is 9.9, which represents a percentage score of 20.1.

Figure 2. Relationship between MDT and MC Item Difficulty (Map Location Questions)

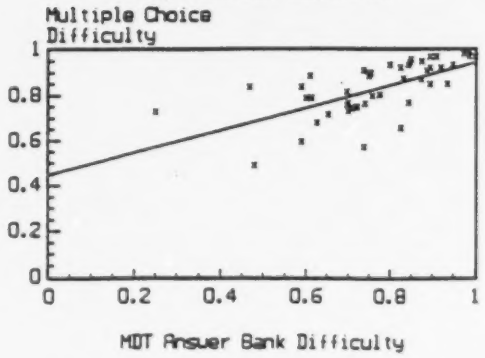
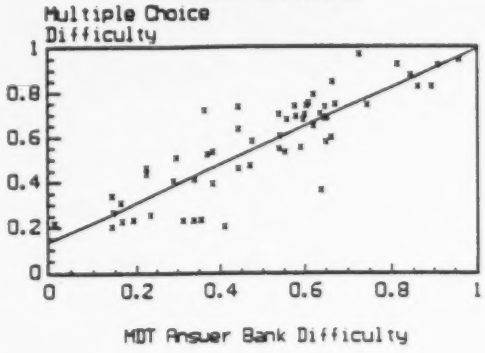


Figure 3. Relationship between MDT and MC Item Difficulty (Concept Questions)



These observations are perhaps more striking when examined as test scores rather than difficulty indices. Using the largest World Geography data set (N = 291), the regression equation predicts that a student missing all MDT questions would answer about seven of the MC equivalents correctly (21% of the total possible). On the other hand, examinees scoring highly on MDT had almost the same high score on MC. The PG90 data set (N = 43) with four-choice MC questions confirms these differences because its regression equation predicts that students missing all MDT questions would answer 28% of the MC correctly.

Regarding item discrimination, Table 2 shows that, although the relationships are positive and statistically significant, the observed MDT and predicted MC values become farther apart as discrimination values increase. MC discrimination is slightly higher than MDT at very low values but the MDT values exceed the MC for the upper part of the range. Thus, when converting questions from MC to MDT, in most cases there would be an improvement in discrimination, and not simply because they become more difficult.

Student "Ability" and Question Style

Because both high and low scores were crucial factors in the above analysis, some of the differences between styles may be a function of student "ability." We recognize that "ability" can include motivation and many other influences and that it is more than the "native"

intelligence as measured by an IQ test. We have currently obtained some "ability" indicators for two of the data sets (WG87 and PG90). We acquired cumulative grade point averages (GPA) for almost all those students, but many (10% and 41%, respectively) did not have A.C.T. or S.A.T. standardized test scores. From the regression analysis, we observed that standardized tests were not good predictors of success with either question style, with either data set, or for GPA. Analyses involving standardized tests will require more data.

Cumulative grade point average (GPA) has a much more pronounced and statistically meaningful effect on scores than standardized tests. Also, variations of GPA have a larger effect

Figure 4A. Relationship between Grade Point Average and MDT Answer Bank Score (Concept Items, Test 4, Spring 1987)

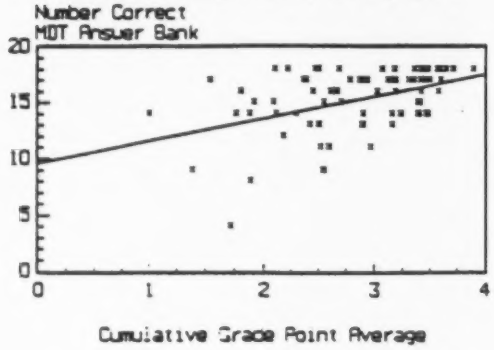


Figure 4B. Relationship between Grade Point Average and MC Score (Concept Items, Test 4, Spring 1987)

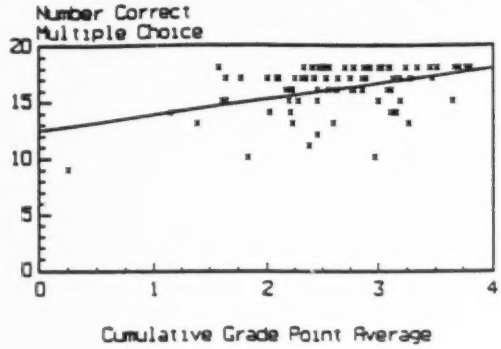


Figure 5A. Relationship between Grade Point Average and MDT Score (Physical Geography, Fall 1990)

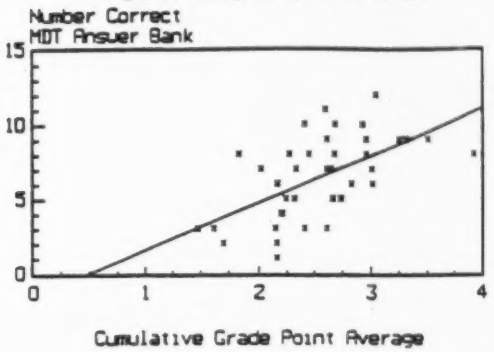
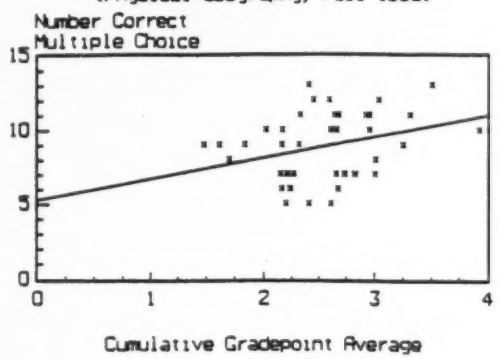


Figure 5B. Relationship between Grade Point Average and MC Score (Physical Geography, Fall 1990)



on MDT than MC scores for parallel questions. MDT scores were less than MC at all grade point levels but the difference between them was much greater at lower GPA than higher GPA (Figs. 4 and 5). For example, the difference between MDT and MC for the map questions of data set WG87 was 12.7 points (out of 100) for a GPA of 1.0, but only 2.6 points at a maximum GPA of 4.0. Assuming most students to have a GPA in the range of 2.0 to 3.0, the difference

in points (9 and 6 points, respectively) between the styles was large relative to the standard 10 point spread in traditional grade categories. The differences were even greater for the PG90 data set (25.8 and 12.5 points at GPA of 2.0 and 3.0, respectively).

CONCLUSIONS

MDT answer bank equivalents of multiple choice are (a) more difficult (averaging 6.4 points less on a 100 point test) and (b) higher in discrimination (0.046, or 14% better). These results are independent of the three categories of questions although "map location" questions are much easier than "concept" questions.

A major conclusion is that test style is very important for lower "ability" examinees but is negligible at higher GPA. The data clearly reveal that the MDT answer bank method can better discriminate among students than can multiple choice, especially when dealing with students at the middle and lower end of the group. This supports the statement of Veloski et al. (1988) that the answer bank method can better identify marginal examinees. Converting multiple choice tests to the MDT answer bank style will lower scores but improve discrimination among examinees, especially for those with lower "ability." We believe that traditional grading scales stressing such thresholds as 90% for an A are detrimental to assessment procedures because they have very little "room" at the top to challenge the best students. Additionally, when used with MC questions, such scales most commonly have relatively lower discrimination power in comparison with MDT answer bank questions. Teachers of geography and other disciplines can effectively and easily employ this method. Relevant materials for several geography courses are available from the authors.

LITERATURE CITED

Anderson, P. S. 1990. Initial experiences with reconsiderative scoring. Paper presented at the conference of the National Council for Measurement in Education, Boston, April.

Anderson, P. S. 1987. The M.D.T. Innovation: Machine Scoring of Fill-in-the-blank tests. Normal, Illinois, Multi-digit Technologies Corporation.

Anderson, P. S., and Hyers, A. D. 1991. Quantitative comparisons of difficulty, discrimination, and reliability of machine-scored completion items and tests (in the MDT un-cued answer bank format) in contrast with statistics from comparable multiple choice questions: the first round of results. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, Illinois, April 4-6.

Brailovsky, C. A., T. Bordage, T. Allen, and H. Dumont. 1988. Writing vs. coding diagnostic impressions in an examination: short-answer vs.long-menu responses. In Research in Medical Education (RIME) Proceedings, the 27th Annual Conference of the Association of American Medical Colleges, Chicago, Illinois, 11-17 November, 201-206.

Veloski, J. J., H. K. Rabinowitz, and M. R. Robeson. 1988. Cueing in multiple choice questions: a reliable, valid and economical solution. In Research in Medical Education (RIME) Proceedings, the 27th Annual Conference of the Association of American Medical Colleges, Chicago, Illinois, 11-17 November, 195-200.